Solutions 1/5

francesco.dotto@uniroma3.it

Exercise 1

Consider the following set of observations:

$$0,\ 43,\ 70,\ 14,\ 22,\ 13,\ 90,\ 50$$

1. Find the mean.
   Answer: the mean is given by
   $$(0 + 43 + 70 + 14 + 22 + 13 + 90 + 50)/8 = 37.75$$

2. Find the median.
   Answer. First, sort the data
   $$0,\ 13{,}14,\ 22{,}43,\ 50,\ 70,\ 90$$
   The median is the average of the two middle values (since the number of observations is even).
   $$(22 + 43)/2 = 32.5$$

Exercise 2

The owner of a company in downtown Rome is concerned about the large use of gasoline by her employees due to urban sprawl, traffic congestion, and the use of energy inefficient vehicles such as SUVs. She'd like to promote the use of public transportation. She decides to investigate how many kilometers her employees travel on public transportation during a typical day. The values for her 10 employees (recorded to the closest kilometer) are

$$0{,}0{,}4{,}0{,}0{,}0{,}10{,}0{,}6{,}0$$

1. Find and interpret the mean, median and mode.
   Answer.
   - Mean
     $$(0 + 0 + 4 + 0 + 0 + 0 + 10 + 0 + 6 + 0)/10 = 2$$
   - Median.
     First sort the data.

0 0 0 0 0 0 0 4 6 10

The median is the average of the two middle values (since the number of observations is even), that is 0.

- Mode:

x_i	n_i
0	7
4	1
6	1
10	1

The mode is 0. Indeed there are 7 zeros in the data set.

2. She has just hired an additional employee. He lives in a different city and travels 90 kilometers a day on public transport. Recompute the mean and median. Describe the effect of this new observation.
Answer.
The mean is given by $(0+0+4+0+0+0+10+0+6+0+90)/11 = 10$.
The median is still 0. (To check, sort the data $0, 0, 0, 0, 0, 0, 0, 4, 6, 10, 90$ and pick the middle value).
Thus, this new observation affects mean but not median.

Exercise 3

The table summarizes responses of 4383 subjects in a recent General Social Survey to the question, *"Whitin the past months, how many people have you known personally that were victims of homicide?"*

Number of Victims	Frequency
0	3944
1	279
2	97
3	40
4 or more	23
Total	**4383**

1. To find the mean, it is necessary to give a score to the "4 or more" class. Find it, using the score 4.5.
Answer

Number of Victims - x_i	Frequency - n_i	$x_i * n_i$
0	3944	0
1	279	279
2	97	194
3	40	120
4.5	23	103.5
Total	**4383**	**696.5**

The mean is given by $696.5/4383 = 0.16$. NOTE that $N = 4383$ (..it is not 5! 5 is the number of distinct values).

2. Find the median. Is the "4 or more" class problematic for it?
Answer. BEAR IN MIND that the frequency distribution is a way to summarize the data; in other words, in this example, you would have 3944 zeros followed by 279 ones, 97 times two, and so on...
The median is the middle score. With 4383 scores, the median is the score in the 2192nd position. Thus, the median is 0. Otherwise, you could use the relative frequencies.

Number of Victims - x_i	Frequency - n_i	$f_i = n_i/N$	F_i
0	3944	0.90	0.90
1	279	0.064	0.96
2	97	0.022	0.986
3	40	0.009	0.995
4 or more	23	0.005	1
Total	**4383**	**1**	

At this point you look at the category corresponding to $F_i = 0.5$ (that is the cumulative relative frequency that splits the data into two equal parts) - that is 0.

3. If 1744 observations shift from 0 to 4 or more, how do the mean and median change?
Answer. The median would still be 0, because there are still 2200 people who gave 0 as a response. The mean would now be $8544.6/4383 = 1.95$.

Number of Victims - x_i	Frequency - n_i	$x_i * n_i$	$f_i = n_i/N$	F_i
0	2200	2200	0.50	0.50
1	279	279	0.064	0.57
2	97	194	0.022	0.588
3	40	120	0.009	0.597
4 or more	1767	7951.5	0.403	1
Total	**4383**	**8544.6**	**1**	

4. Why is the median the same for parts 2 and 3, even though the data are so different?
Answer. The median is the same for both because the median ignores much of the data. The data are highly discrete; hence, a high proportion of the data falls at only one or two values. The mean is better in this case because it uses the numerical values of all of the observations, not just the ordering.

Exercise 4

Consider the following two sets of observations:

Set 1: 2,3,3,3,4,4,4
Set 2: 2,3,3,3,3,3,4

1. Find the variance for each data set.
   Answer.

   - Set 1: Mean=3.29

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	-1.2857143	1.65306122
3	-0.2857143	0.08163265
3	-0.2857143	0.08163265
3	-0.2857143	0.08163265
4	0.7142857	0.51020408
4	0.7142857	0.51020408
4	0.7142857	0.51020408
Total	0	3.29

   The variance is given by $3.29/7 = 0.49$.

   - Set 2: Mean=3.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	-1	1
3	0	0
3	0	0
3	0	0
3	0	0
3	0	0
4	1	1
Total	0	2

   The variance is given by $2/7 = 0.29$.

2. Which data set shows more variability?
   Answer. Set 1.

Exercise 5

A company decides to investigate the amount of sick leave taken by its employees. A sample of 8 employees yields the following numbers of days of sick leave taken in the past year

0,0,4,0,0,0,6,0

1. Find and interpret the range.
   Answer. Range=Max-Min=6-0=6

2. Find and interpret the standard deviation s.
   Answer.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0	-1.25	1.5625
0	-1.25	1.5625
4	2.75	7.5625
0	-1.25	1.5625
0	-1.25	1.5625
0	-1.25	1.5625
6	4.75	22.5625
0	-1.25	1.5625
Total	0	39.5

   The standard deviation is given by $\sqrt{39.5/8} = 2.22$ The deviation from the mean that we *generally* have is 2.22.

3. Suppose that 6 was incorrectly recorded and is supposed to be 60. Redo parts 1 and 2 with the correct data and describe the effect of this outlier.
   Answer.

   Now the range is 60.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0	-8	64
0	-8	64
4	-4	16
0	-8	64
0	-8	64
0	-8	64
60	52	2704
0	-8	64
Total	0	3104

   The standard deviation is given by $\sqrt{3104/8} = 19.70$
   The range and mean both increase when an outlier is added.

Exercise 6

The mean and standard deviation of a sample may change if data are rescaled.

1. Scores on a difficult exam have a mean of 57 and a standard deviation of 20. The teacher boosts all the scores by 20 points before awarding grades. Report the mean and the variance of the boosted scores.
   Answer
   mean=77; s=20.
   $\bar{y} = a + b\bar{x}$. $a = 20$ and $b = 1$

2. Referring to point 1, what happens to the mean if the students get a grade rise of 3%?
   Answer
   mean=58.71; s=0.6;
   $\bar{y} = a + b\bar{x}$. $a = 0$ and $b = 1.03$
   Thus $s^2 = b^2 \times 20^2 = 20.6^2 = 424.36$, while $\bar{y} = 58.71$

3. Suppose that the annual income for some group has a mean of $ 39,000 and a standard deviation of $ 15,000. Values are converted to euros. If one euro equals $2.00, report the mean and standard deviation in European currency.
   Answer
   mean=19500; s=7500; $\bar{y} = a + b\bar{x}$. $a = 0$ and $b = 0.5$.
   The mean of the annual income converted to euros is 19500; while the standard deviation converted to euros is 7500.

Exercise 7

A professor examined the results of the first exam given in her statistics class. The scores were

$$70,\ 84,\ 59,\ 73,\ 86,\ 35,\ 81,\ 75.$$

1. Find the mean and the median.
   Answer
   The mean is given by

   $$(70 + 84 + 59 + 73 + 86 + 35 + 81 + 75)/8 = 70.375;$$

   while to find the median, first you have to sort the data,

   $$35, 59, 70, 73, 75, 81, 84, 86.$$

   It follows that the median is given by the average of the two middle values, (73+75)/2 = 74

2. Would you guess that the distribution is skewed or roughly symmetric? Why?
   Answer
   Since the median is greater than the mean, the distribution is skewed to the left.

3. Find the standard deviation.
   Answer

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
35	-35.375	1251.39
59	-11.375	19.39
70	-0.375	0.14
73	2.625	6.89
75	4.625	21.39
81	10.625	112.89
84	13.625	185.64
86	15.625	244.14
Total	0	1951.875

   The variance is given by $1951.875/8 = 243.9844$. Thus, $s = \sqrt{243.9844} = 15.62$.

Exercise 8

For the question "How many children have you ever had?", the results were

No.Children	0	1	2	3	4
Count	25	15	20	5	0

1. Find the variance and the standard deviation.
   Answer

x_i	n_i	$n_i * x_i$	$(x_i - \bar{x})^2$	$n_i * (x_i - \bar{x})^2$
0	25	0	$(-1.08)^2$	29.16
1	15	15	$(-0.08)^2$	0.096
2	20	40	$(0.92)^2$	16.928
3	5	15	$(1.92)^2$	18.432
4	0	0	$(2.92)^2$	0
Total	65	70		64.616

   To compute the variance and the standard deviation, first we need to compute the mean, $\bar{x} = 70/65$. The variance is obtained as $s^2 = 64.616/65 = 0.99$; thus, $s = \sqrt{0.99} = 0.995$.

Exercise 9

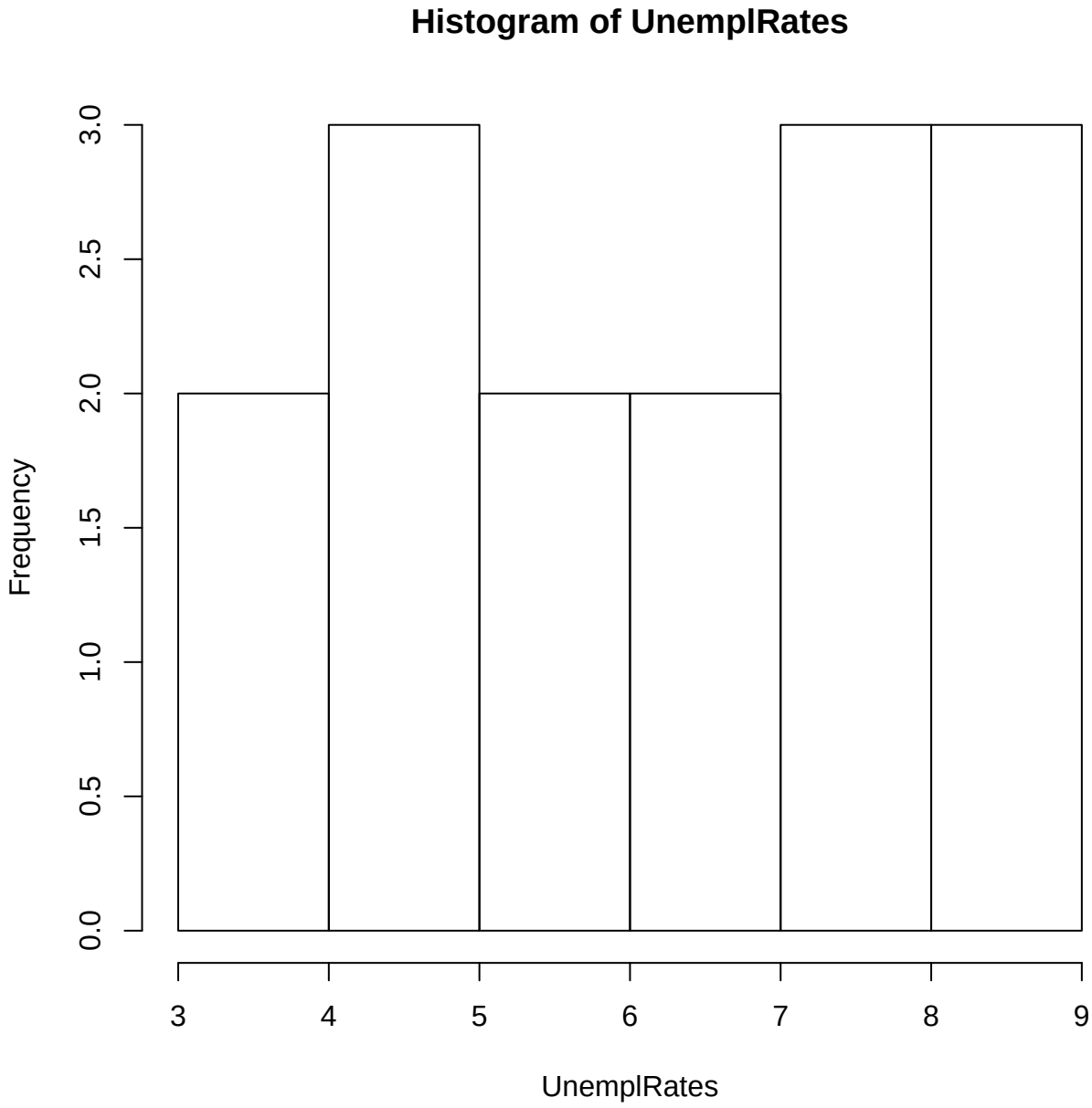
The data values below represent the prices per share of the 20 most actively traded stocks on the New York Stock Exchange (rounded to the nearest dollar) on February 18 2011.

5,15,2,16,5,5,21,33,19,9,7,9,48,39,52,17,85,13,35,10

1. Construct a histogram.
   Answer
   First we need to construct the frequency table

Interval	Frequency
1 to 10	8
11 to 20	5
21 to 30	1
31 to 40	3
41 to 50	1
51 to 60	1
61 to 70	0
71 to 80	0
81 to 90	1

   The corresponding histogram is depicted in the following figure.



2. Find the median, the first quartile, and the third quartile.
   Answer
   Median=(15+16)/2=15.5. $Q_1 = (7+9)/2 = 8$. $Q_3 = (33+35)/2 = 34$.
   SUMMARY: Finding Quartiles

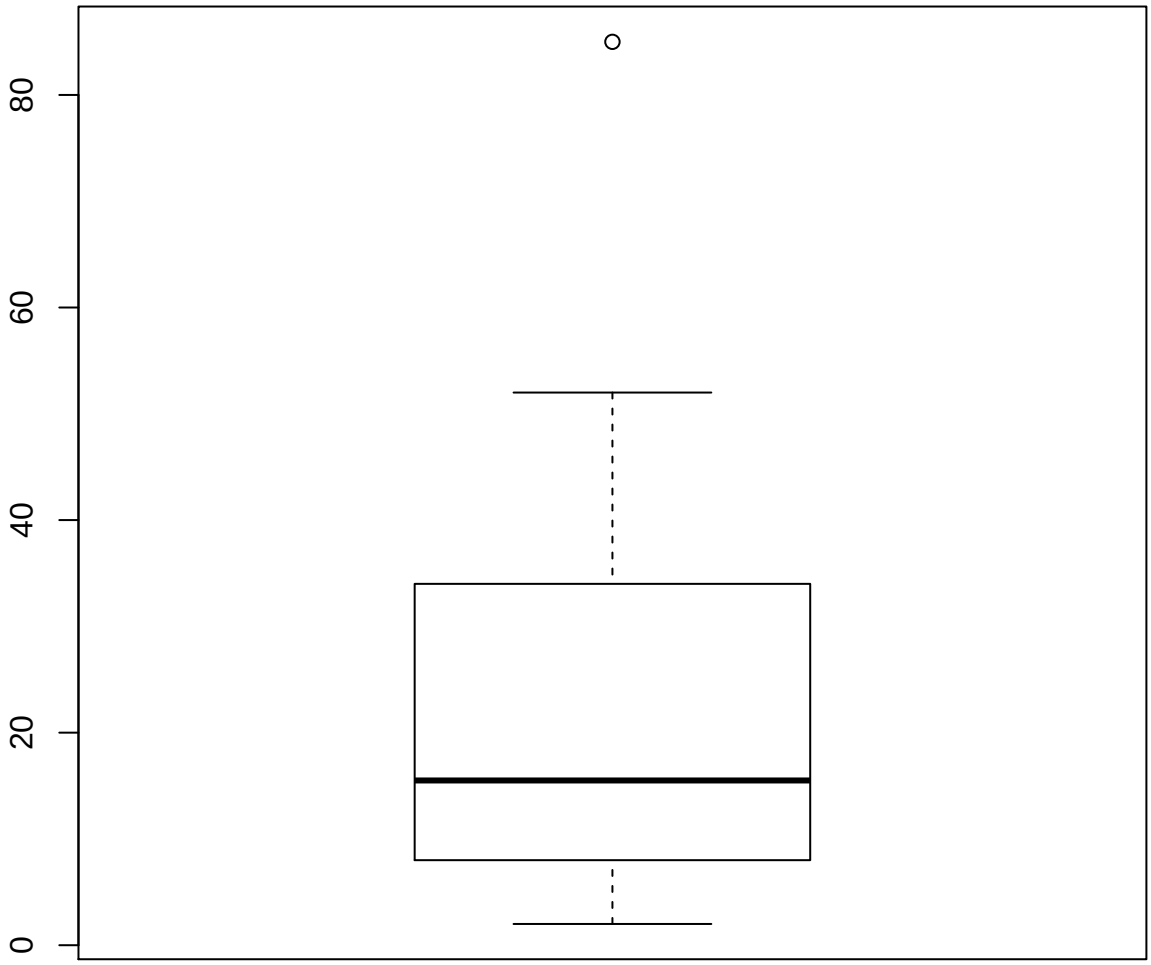
   (a) Arrange the data in order.

   (b) Consider the median. This is the second quartile, Q_2.

   (c) Consider the lower half of the observations (excluding the median itself if n is odd). The median of these observations is the first quartile, Q_1.

(d) Consider the upper half of the observations (excluding the median itself if n is odd). Their median is the third quartile, Q_3.

3. Sketch a box plot. What feature of the distribution displayed in the histogram is not obvious in the box plot? (Hint: Are there any gaps in the data?)
Answer



The box plot does show that the observation of 85 is a potential outlier, as confirmed

by the following formula $1.5 \times (Q_3 - Q_1) + Q_3 = 1.5 \times (34 - 8) + 34 = 73$. Indeed 85 is greater than 73. However, the boxplot doesn't show the other gaps in the distribution.

Exercise 10

The fertility rate for a nation is measured as the average number of children per adult woman. The table below shows results for western European nations, the United States, Canada, and Mexico, as reported by the United Nations in 2005.

Country	Fertility	Country	Fertility
Austria	1.4	Netherlands	1.7
Belgium	1.7	Norway	1.8
Denmark	1.8	Spain	1.3
Finland	1.7	Sweden	1.6
France	1.9	Switzerland	1.4
Germany	1.3	United Kingdom	1.7
Greece	1.3	United States	2.0
Ireland	1.9	Canada	1.5
Italy	1.3	Mexico	2.4

1. Find the quartiles (Q_1, Q_2, Q_3) for the fertility rates.
   Answer
   First sort the data

   $$1.3, 1.3, 1.3, 1.3, 1.4, 1.4, 1.5, 1.6, 1.7, 1.7, 1.7, 1.7, 1.8, 1.8, 1.9, 1.9, 2.0, 2.4$$

   $Q_1 = 1.4$. $Q_2 = 1.7$. $Q_3 = 1.8$

2. Find the interquantile range (IQR).
   Answer
   IQR=1.8-1.4=0.4

3. Find the five-number summary. **Answer**
   Min=1.3, $Q_1 = 1.4$, $Q_2 = 1.7$, $Q_3 = 1.8$, Max=2.4.

Exercise 11

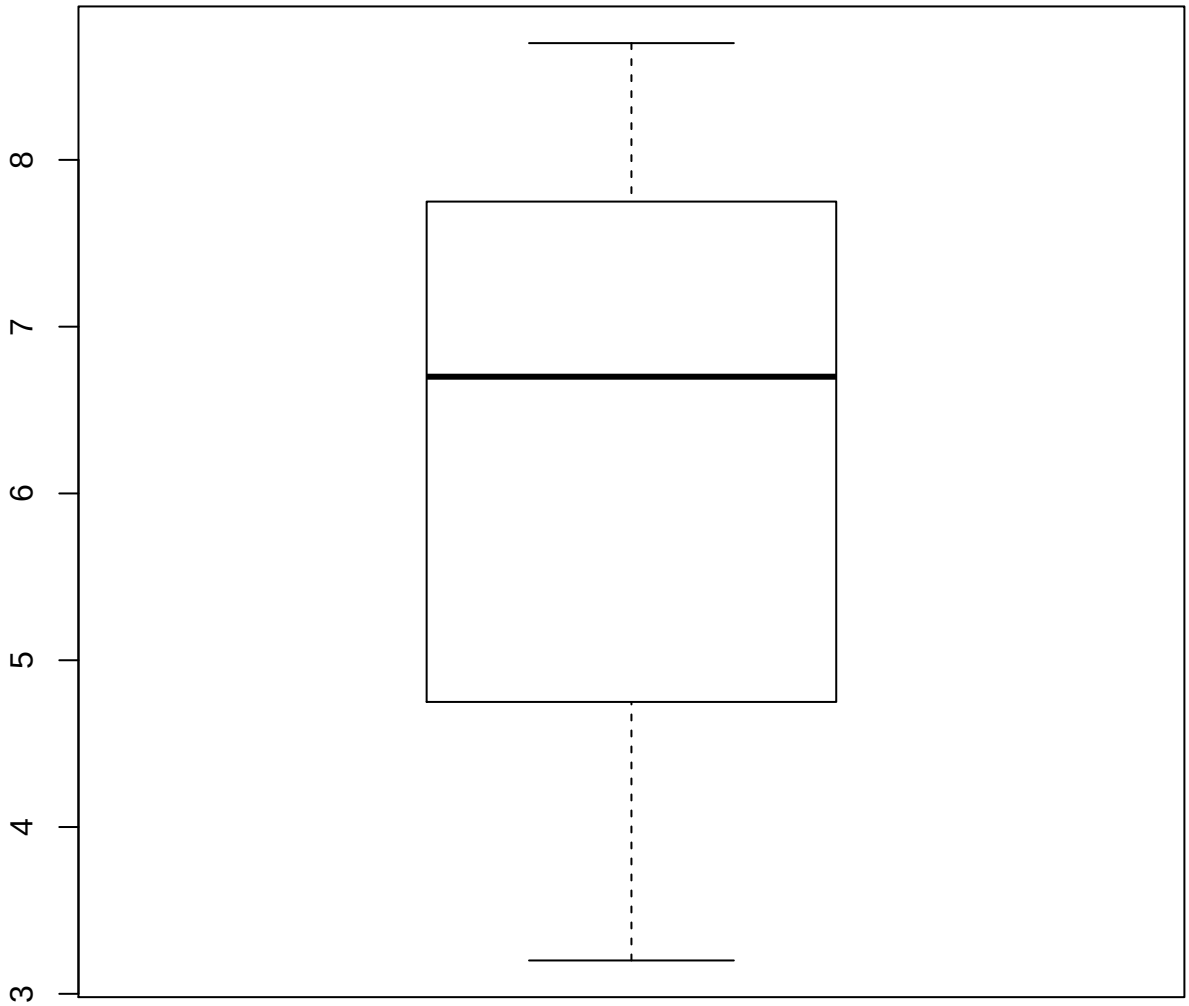
The 2007 unemployment rates of countries in the European Union are shown in the table below.

Country	Unemployment rate	Country	Unemployment rate	Country	Unemployment rate
Belgium	7.8	France	8.4	Italy	6.7
Denmark	3.2	Portugal	7.2	Finland	7.0
Germany	7.7	Netherlands	3.6	Austria	4.5
Greece	8.7	Luxembourg	5.0	Sweden	6.0
Spain	8.6	Ireland	4.4	U.K.	5.4

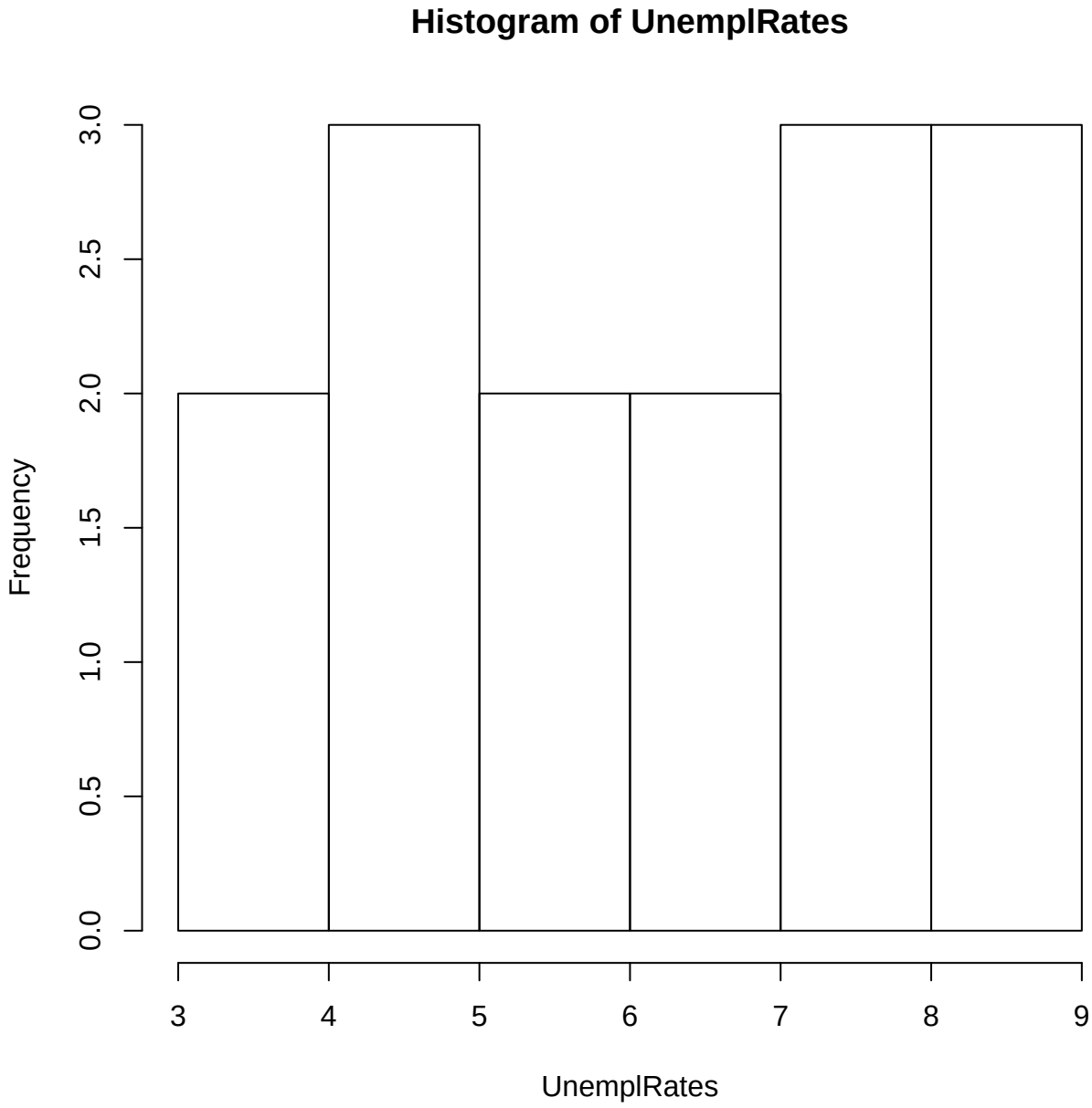
1. Identify the five-number summary, and sketch a box plot.
   Answer
   First sort the data

$$3.2, 3.6, 4.4, 4.5, 5, 5.4, 6, 6.7, 7, 7.2, 7.7, 7.8, 8.4, 8.6, 8.7$$

   Min=3.2, $Q_1 = 4.5$, $Q_2 = 6.7$, $Q_3 = 7.8$, Max=8.7.



2. Provide a graphical representation of the distribution.
   Answer



3. Greece had the highest unemployment rate of 8.7. Is it an outlier? Explain.
   Answer
   According to the following formula $1.5 \times (Q_3 - Q_1) + Q_3 = 1.5 \times (7.8 - 4.5) + 7.8 = 12.25$, it is not a potential outlier.

4. Find the mean and standard deviation.
   Answer
   The mean is 6.28; while the standard deviation is 1.84.

5. What unemployment value for a country would have a z-score equal to 0?
   Answer

A z-score of 0 indicates that the country's unemployment rate is zero standard deviations from the mean; hence, the unemployment rate is equal to the mean. In this case, a country with an unemployment rate of 6.3 would have a z-score of 0.

Exercise 12

Looking at the following table

Gender	Binge Drinker	Non-Binge Drinker	Total
Male	1908	2017	**3925**
Female	2854	4125	**6979**
Total	**4762**	**6142**	**10904**

1. Identify the response variable and the explanatory variable.
   Answer
   Response: binge drinking. Explanatory: gender

2. Report the cell counts of subjects who were (i) male and a binge drinker, (ii) female and a non-binge drinker.
   Answer
   (i) 1908; (ii) 4125

3. Construct a contingency table that shows the conditional proportions of sampled subjects who do or do not binge drink, given gender.
   Answer

Gender	Binge Drinker	Non-Binge Drinker	Total
Male	0.49	0.51	**1**
Female	0.41	0.59	**1**

   While the unconditional proportions of sampled subjects who do or do not binge drinks are given by

Gender	Binge Drinker	Non-Binge Drinker	Total
Male	0.49	0.51	**1**
Female	0.41	0.59	**1**
Total	**0.44**	**0.56**	

4. Based on part 3, does it seem that there is an association between binge drinking and gender?
   Answer
   It appear that men are more likely than women to be binge drinkers. Indeed looking at the unconditional proportions we expect to see 0.44 binge drinkers and 0.56 non-binge drinkers. Focusing on the binge drinkers, the men are more likely to be binge drinkers

than expected (0.49 compared to 0.44), while the women are less binge drinkers than expected. (0.41 compared yo 0.44). Conversely, looking at the non-binge drinkers, the women are more non-binge drinkers than expected (0.59 rather than 0.56), while the men are less non-binge drinkers than expected (0.51 compared to 0.56).

Exercise 13

The table shows results of whether the death penalty was imposed in murder trials in Florida between 1976 and 1987. For instance, the death penalty was given in 53 out 467 cases in which a white defendant had a white victim.

Victim's Race	Defendant's Race	YES Death Penalty	NO Death Penalty	Total
White	White	53	414	**467**
White	Black	11	37	**48**
Black	White	0	16	**16**
Black	Black	4	139	**143**

1. Consider only the cases in which the victim was white. Find the conditional proportions that got the death penalty when the defendant was white and when the defendant was black. Describe the association.
   Answer

	YES Death Penalty	NO Death Penalty
White defendant	0.11	0.89
Black defendant	0.23	0.77
	0.12	0.87

   Black defendant were more likely than were white defendant to get the death penalty when the victim was white.

2. Repeat part 1 for cases in which the victim was black. **Answer**

	YES Death Penalty	NO Death Penalty
White defendant	0	1
Black defendant	0.03	0.97
	0.03	0.97

   Black defendant were more likely than were white defendant to get the death penalty when the victim was black.
   Answer

	YES Death Penalty	NO Death Penalty	Total
White defendant	53	430	**483**
Black defendant	15	176	**191**
Total	**68**	**606**	**674**

3. Construct a summary contingency table that describes the association between the death penalty verdict and defendant's race, ignoring the information about the victim's race.

4. Find the conditional proportions and describe the association.

   Answer

	YES Death Penalty	NO Death Penalty
White defendant	0.11	0.89
Black defendant	0.08	0.92
	0.10	0.90

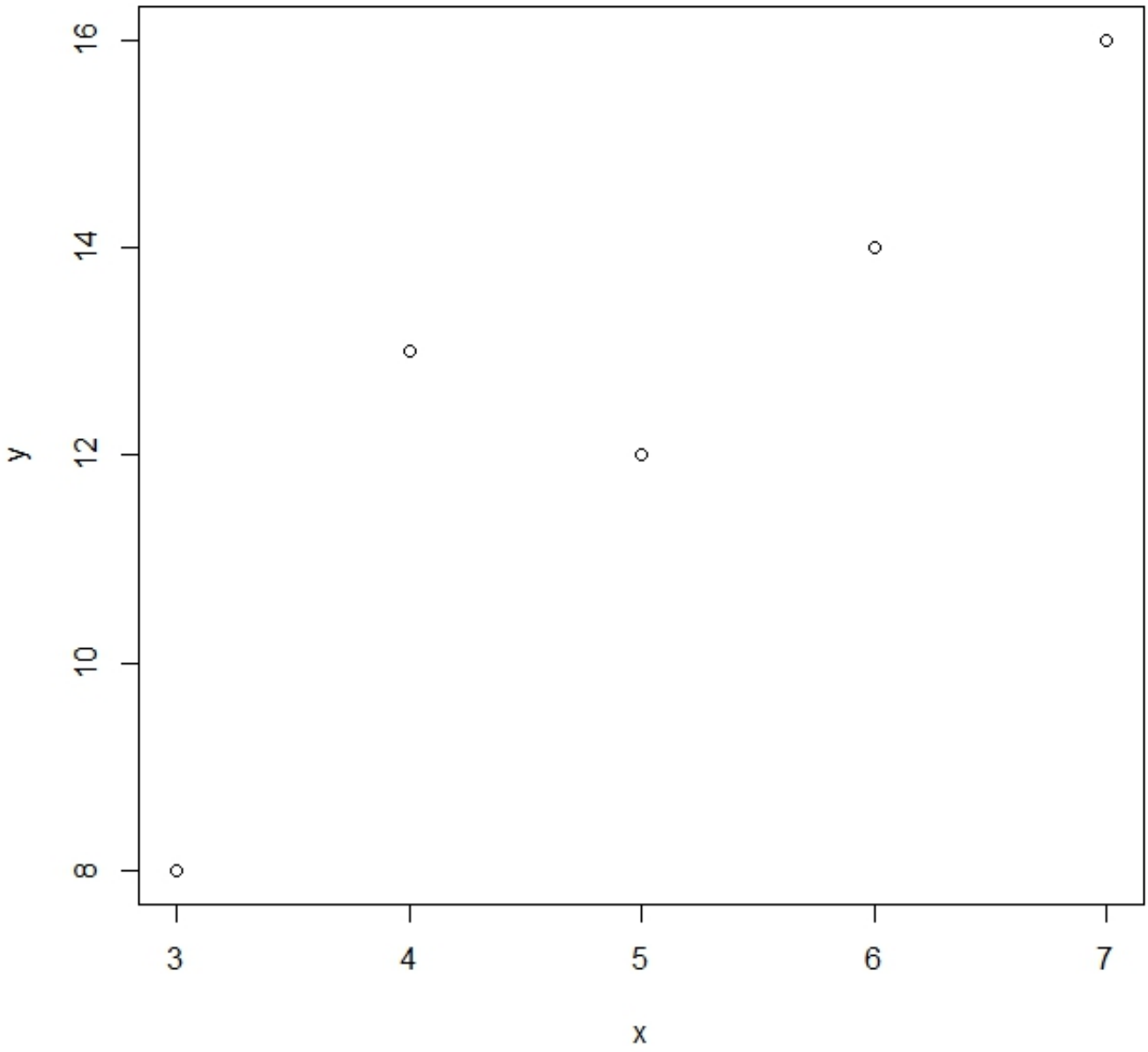
   These data indicate that white defendant were more likely than were black defendants to get the death penalty.
   NOTE: The association changes when the race of the victim is included.

Exercise 14

Consider the data

x	y
3	8
4	13
5	12
6	14
7	16

1. Sketch a scatterplot.
   Answer



2. Would you expect a positive association, a negative association or no association between x and y?
   Answer
   Looking at the points we expect to have a positive association between x and y. Indeed they could be interpolated by a positive line.

3. Compute the correlation coefficient, r.
   Answer

$\bar{x} = 5$, $\bar{y} = 12.6$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) \times (y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
3	8	-2	-4.6	9.2	4	21.16
4	13	-1	0.4	-0.4	1	0.16
5	12	0	-0.6	0	0	0.36
6	14	1	1.4	1.4	1	1.96
7	16	2	3.4	6.8	4	11.56
				17	**10**	**35.2**

It follows that $r = 17/\sqrt{(10 \times 35.2)} = 0.91$.

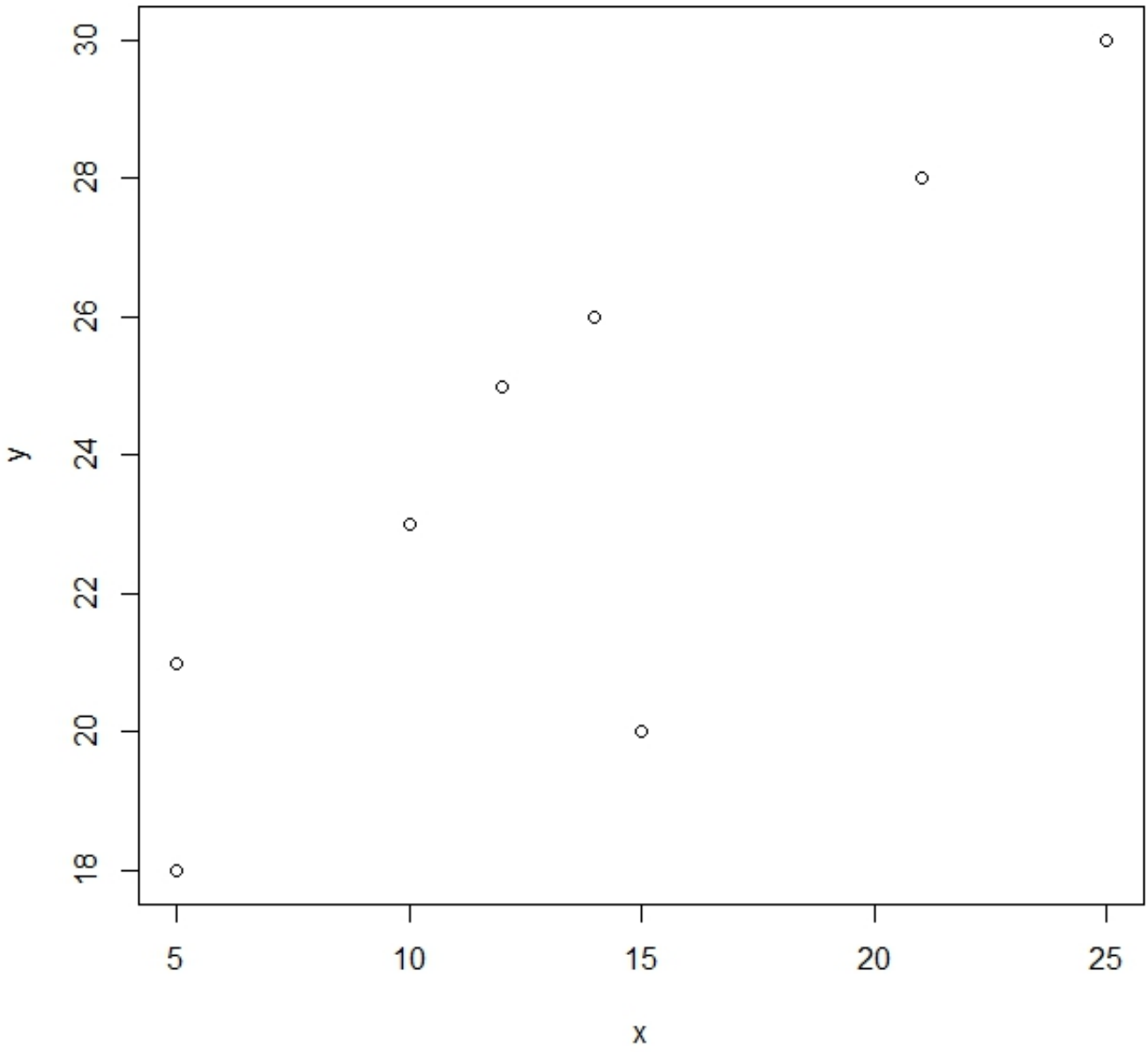
Exercise 15

An instructor of Statistics collected data from one of her classes in Spring 2016 to investigate the relationship between Study time per week (number of hours) to predict the final grade. For the 8 students in her class the data were as shown in the table.

Student	**Study Time**	**Grade**
1	14	26
2	25	30
3	15	20
4	5	18
5	10	23
6	12	25
7	5	21
8	21	28

1. Identify the response variable and the explanatory variable.
   Answer
   Response variable: Grade. Explanatory variable: Study Time.

2. Construct a scatterplot.
   Answer



3. Find and interpret the correlation.
   Answer
   $\bar{x} = 13.375$, $\bar{y} = 23.875$

Student	**Study Time**	**Grade**	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) \times (y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
1	14	26	0.62	2.12	1.31	0.38	4.49
2	25	30	11.62	6.12	71.11	135.02	37.45
3	15	20	1.62	-3.88	-6.29	2.62	15.05
4	5	18	-8.38	-5.88	49.27	70.22	34.57
5	10	23	-3.38	-0.88	2.97	11.42	0.77
6	12	25	-1.38	1.12	-1.55	1.90	1.25
7	5	21	-8.38	-2.88	24.13	70.22	8.29
8	21	28	7.62	4.12	31.39	58.06	16.97
					172.34	**349.84**	**118.84**

It follows that $r = 172.34/\sqrt{(349.84 \times 118.84)} = 0.8452$. It means that Study Time and Grade are highly positive correlated. When the Study Time increases, the Grade increases too.